A Spelling Suggestion Technique for Terminology Servers

Guy Divita, Allen C. Browne, Tony Tse, May L. Cheh, Russell F. Loane Ph.D, Myriam Abramson
Lister Hill National Center for Biomedical Communications,
National Library of Medicine
Bethesda, Maryland

INTRODUCTION

The Terminology Server project at NLM is concerned with bridging the gap between user's terminology and that of the information systems they wish to access. Examination of user logs of the NLM home page showed that many unproductive queries resulted from spelling errors. In response to this problem we hope to add a spelling correction capability to the UMLS lexical tools and our terminology server. The projected component will suggest correct spellings for potentially misspelled words or terms. The experiment reported here is an evaluation of our effort to develop an efficient spelling suggestion algorithm.

METHODS

We used a combination of algorithms, and a cache of correct and incorrect spelling pairs to select a set of candidate suggestions and another algorithm to order this set for presentation to users. We combined a variety of the common spelling algorithms to cover different types of spelling errors. We chose Metaphone to cover spelling errors due to phonetic spellings. We used a character based bi-gram approach to cover OCR errors, typographic errors and other errors due to addition, deletion, substitution or transposition of letters. Each algorithm selects a set of candidate correct spellings from the dictionary for each term. The candidates are ranked by minimum edit distance.

The goal was to match or exceed the performance of spelling correctors developed elsewhere. We compared our algorithm with Aspell whose own evaluation compares it to Ispell and Microsoft Word. First we evaluated both systems using the full set of inflected words from the SPECIALIST lexicon. We then used the normal dictionary that came with Aspell. We based the tests on two collections of misspellings and corresponding correct spellings. We used the set that came with Aspell, and another we developed specific to the medical domain.

RESULTS

In all cases, using the three combinations of dictionaries and collections of misspelled words (See Table 1) both systems are comparable with respect to finding the correct spelling as the first suggestion or in the top five suggestions. Our algorithm performed statistically significantly better than Aspell in one case: using Aspell misspelled words with the SPECIALIST dictionary and finding the correct spelling in the top five suggestions. In no case was our algorithm statistically significantly worse.

Misspelled Words		Dictionary			
		SPECIALIST		Aspell (Normal)	
	Hits:	1	5	1	5
Aspell	Ours	**263**	**385**	260	393
n = 546	Aspell*	263	356	259	410
MEDLINE*plus*	Ours	144	188		
n = 245	Aspell*	149	179		

Table 1: Number of matches found by our algorithm and Aspell

CONCLUSION

We have developed a viable component to the lexical tools for handling spelling issues. This technique, in conjunction with the suite of lexical tools provides a powerful and flexible set to apply to terminology servers and retrieval systems.

FUTURE DIRECTIONS

We are experimenting with an algorithm that uses an edit distance that breaks the ties with the prior probability of the word computed as a frequency count on a corpus. We are also experimenting with an algorithm that uses a weighted edit distance computing the cost of an operation as the conditional probability of a letter given a word in the dictionary. The minimum cost represents the maximum probability of obtaining the word through the transcript operations. We plan to examine methods to create target-specific dictionaries for spelling suggestion. The method can be expanded to index and retrieve multi-word terms, a feature not found in other spell checkers including Aspell.